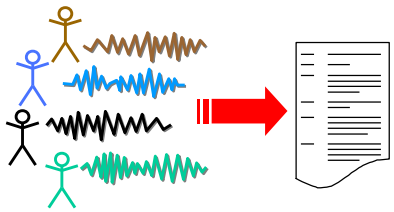
Who Spoke When

Speaker-ID and Speaker-Type Metadata Diarization



May 20, 2003

Participants

- CTS
  - Cambridge
  - ISL
  - Lincoln
- BNews
  - Cambridge
  - CLIPS
  - ELISA (CLIPS + LIA)
  - ICSI
  - LIA
  - LIMSI
  - Lincoln
  - Panasonic

Data

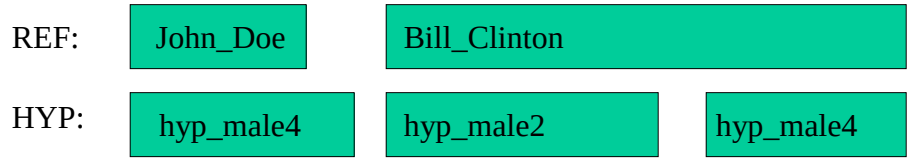
- CTS
  - 36 calls
  - Half are Switchboard-cellular and half Fisher
  - Balanced by gender
    - 11 male ↔ male
    - 14 female ↔ male
    - 11 female ↔ female
  - two-channel data
    - one speaker per channel
    - no speaker error
- BNews
  - 3 shows
    - All from February 2001
    - The first three shows (in time) from the STT set
      - VOA_ENG
      - PRI_TWD
      - MNB_NBW
  - Total of 57 speakers
    - 45 male
    - 12 female

Test Conditions

- Who Spoke When -- diarization by speaker-ID
- What Speaker-type Spoke When (use type as ID)
  - adult_male
  - adult_female
  - child
  - unknown
- Two scorings: time-based and word-based
  - Word-based scoring omits all False-alarm errors

Types of Diarization Error

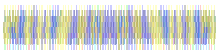
We map hyp spkrs to ref spkrs to maximize the mapped time.
If John_Doe mapped to hyp_male4 and Bill Clinton mapped to hyp_male2, then we get...



ERRORS:

False Alarm 

Miss 

Speaker Err 

DiarizationError = FalseAlarm + Miss + SpeakerErr

$$Error_{SpkrSeg} = \frac{\sum_{\substack{\text{all}\\ \text{segs}}} \{\text{dur}(seg) \cdot (\max(N_{Ref}(seg), N_{Sys}(seg)) - N_{Correct}(seg))\}}{\sum_{\substack{\text{all}\\ \text{segs}}} \{\text{dur}(seg) \cdot N_{Ref}(seg)\}}$$

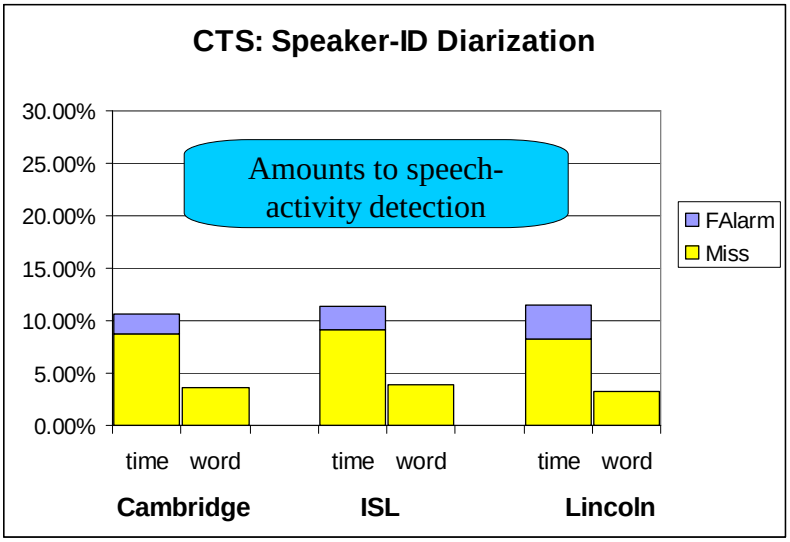
where the speech data file is divided into contiguous segments at all speaker change points and where, for each segment, *seg*:

$\text{dur}(seg)$ = the duration of *seg*,
$N_{Ref}(seg)$ = the # of reference speakers speaking in *seg*,
$N_{Sys}(seg)$ = the # of system speakers speaking in *seg*,
$N_{Correct}(seg)$ = the # of reference speakers speaking in *seg* for whom their matching (mapped) system speakers are also speaking in *seg*.

The numerator is diarization error time.

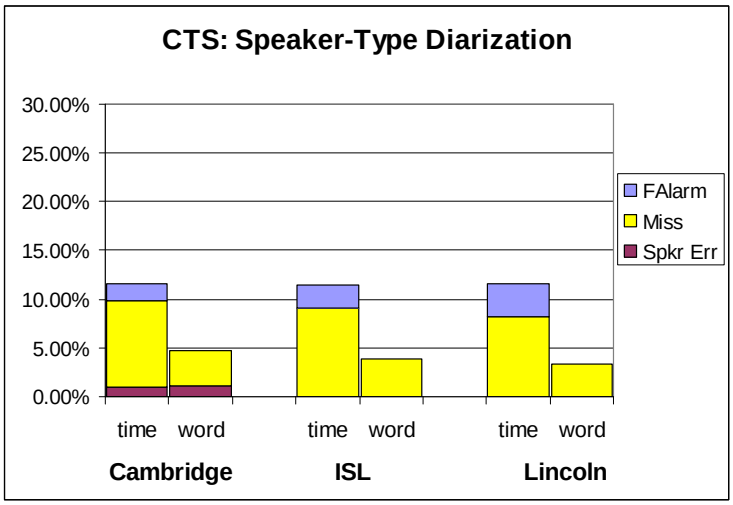
The denominator is speaker time

CTS: Who Spoke When

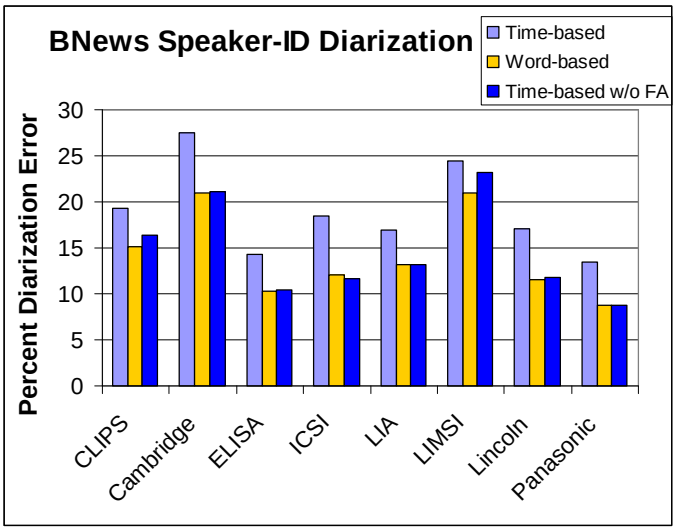


CTS: What Speaker-type Spoke When

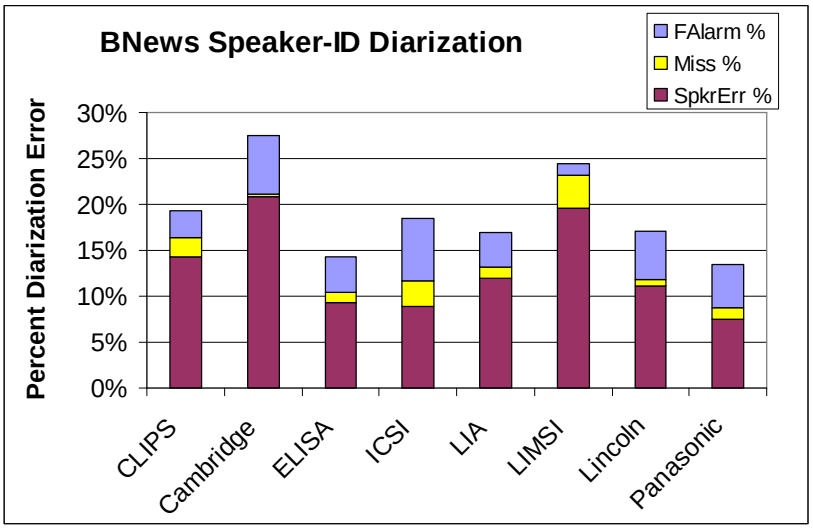
(adult_male / adult_female / child / unknown)



BNews: Who Spoke When

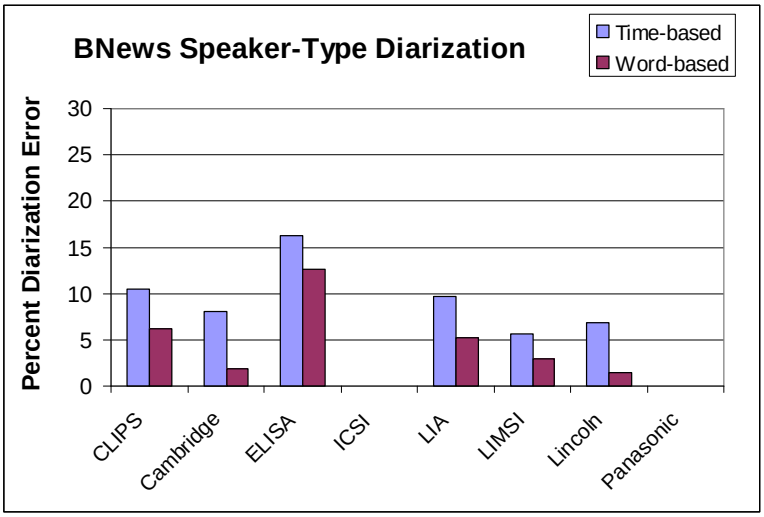


BNews: time-based Who Spoke When (showing the three error components)

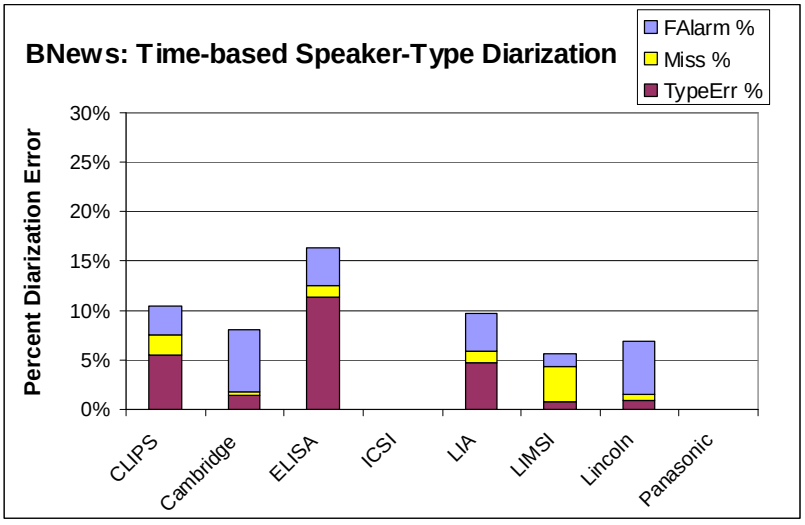


BNews: What Speaker-type Spoke When

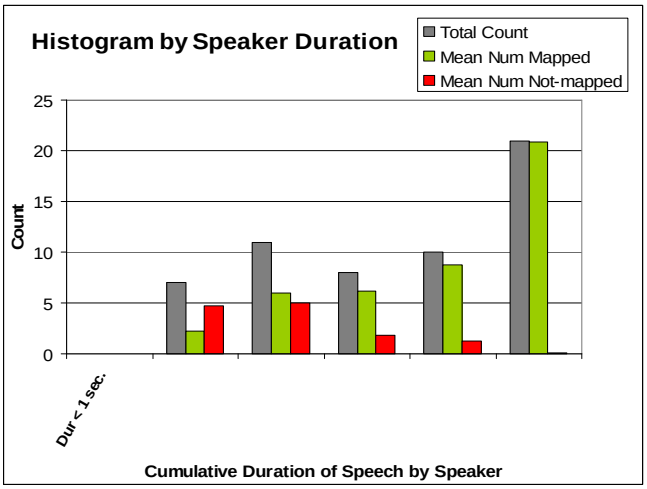
(adult_male / adult_female / child / unknown)



BNews: What Speaker-type Spoke When (showing the three error components)

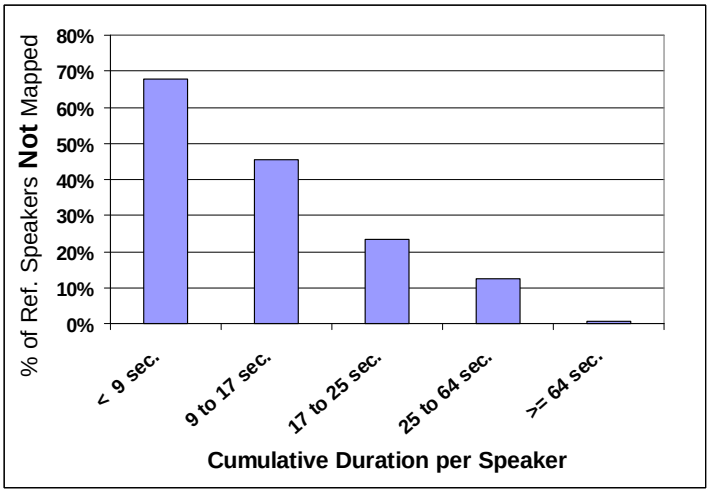


Amount of Speech per Speaker



BNews: Fraction of Speakers Not Mapped as a function of the Speakers' Duration

In essence, % of speakers lumped with other(s)



Fraction of BNews Speakers Mapped

- On average, 74% of Male speakers were mapped
- On average, 92% of Female speakers mapped

- But the difference appears to be accounted for by the cumulative duration of speech by the speakers (cf. bar-chart in previous slide)
  - Eight of the twelve female speakers in this test set spoke more than 32 seconds (six spoke > 64 seconds), so the speaker-ID diarization results were good for female BNews speakers.

Summary

- CTS
  - Results dominated by speech activity detection
  - Time-based diarization error similar for all sites
    - about 8 to 9% Miss, plus 2 to 3% FA

- BNews
  - Speakers who speak > 25 seconds likely to be mapped
  - Word-based BNews results roughly equivalent to time-based results on just Speaker-error plus Miss-error
  - Word-based metric using STT output could be more interesting than current versions using just ref words

Tables of Results: CTS by Speaker-ID

CTS					
Time based, Speaker-ID based					Diarization
		Miss	Spkr Err	FAlarm	Err
	Cambridge	8.78%	0.00%	1.86%	10.65%
	ISL	9.07%	0.01%	2.34%	11.42%
	Lincoln	8.24%	0.00%	3.28%	11.52%
CTS					
Word based, Speaker-ID based					Diarization
		Miss	Spkr Err	FAlarm	Err
	Cambridge	3.67%	0.01%	0.00%	3.68%
	ISL	3.90%	0.01%	0.00%	3.91%
	Lincoln	3.31%	0.00%	0.00%	3.31%

Tables of Results: CTS by Speaker-Type

CTS					
Time based, Speaker-type based					Diarization
		Miss	SpkrType	FAlarm	Err
	Cambridge	8.78%	0.99%	1.86%	11.63%
	ISL	9.07%	0.00%	2.34%	11.41%
	Lincoln	8.24%	0.00%	3.28%	11.52%
CTS					
Word based, Speaker-type based					Diarization
		Miss	SpkrType	FAlarm	Err
	Cambridge	3.67%	1.09%	0.00%	4.77%
	ISL	3.90%	0.00%	0.00%	3.90%
	Lincoln	3.31%	0.00%	0.00%	3.31%

Tables of Results: BNews by Speaker-ID

BNews					
Time-based, Speaker-ID based					Diarization
		Miss	Spkr Err	FAlarm	Err
	CLIPS	1.99%	14.33%	2.93%	19.25%
	Cambridge	0.37%	20.78%	6.30%	27.44%
	ELISA	1.15%	9.32%	3.77%	14.24%
	ICSI	2.87%	8.85%	6.81%	18.52%
	LIA	1.15%	11.98%	3.77%	16.90%
	LIMSI	3.56%	19.64%	1.27%	24.47%
	Lincoln	0.72%	11.06%	5.29%	17.07%
	Panasonic	1.28%	7.49%	4.68%	13.44%
BNews					
Word based, Speaker-ID based					Diarization
		Miss	Spkr Err	FAlarm	Err
	CLIPS	1.04%	14.08%	0.00%	15.11%
	Cambridge	0.35%	20.59%	0.00%	20.94%
	ELISA	0.66%	9.68%	0.00%	10.34%
	ICSI	3.09%	8.95%	0.00%	12.03%
	LIA	0.66%	12.54%	0.00%	13.20%
	LIMSI	1.95%	18.97%	0.00%	20.92%
	Lincoln	0.39%	11.10%	0.00%	11.49%
	Panasonic	0.98%	7.78%	0.00%	8.76%

Tables of Results: BNews by Speaker-type

BNews					
Time based, Speaker-type based					Diarization
		Miss	SpkrType	FAlarm	Err
	CLIPS	1.99%	5.54%	2.93%	10.46%
	Cambridge	0.37%	1.36%	6.30%	8.02%
	ELISA	1.15%	11.38%	3.77%	16.30%
	ICSI				
	LIA	1.15%	4.76%	3.77%	9.68%
	LIMSI	3.56%	0.80%	1.27%	5.63%
	Lincoln	0.72%	0.84%	5.29%	6.84%
	Panasonic				
BNews					
Word based, Speaker-type based					Diarization
		Miss	SpkrType	FAlarm	Err
	CLIPS	1.04%	5.17%	0.00%	6.21%
	Cambridge	0.35%	1.51%	0.00%	1.86%
	ELISA	0.66%	12.03%	0.00%	12.68%
	ICSI	-----	-----	-----	-----
	LIA	0.66%	4.61%	0.00%	5.27%
	LIMSI	1.95%	1.03%	0.00%	2.98%
	Lincoln	0.39%	1.05%	0.00%	1.44%
	Panasonic	-----	-----	-----	-----